When is Menzerath-Altmann law mathematically trivial?

Ramon Ferrer-i-Cancho[a,*], Jaume Baixeries[a], Antoni Hernández-Fernández[a,b]

[a] *Complexity and Quantitative Linguistics Lab. Departament de Llenguatges i Sistemes Informàtics. TALP Research Center/LARCA. Universitat Politècnica de Catalunya. Barcelona (Catalonia), Spain.*

[b] *Departament de Lingstica General. Universitat de Barcelona. Barcelona (Catalonia), Spain.*

Abstract

Menzerath's law, the tendency of Z, the mean size of the parts, to decrease as X, the number of parts, increases is found in language, music and genomes. Recently, it has been argued that the presence of the law in genomes is an inevitable consequence of the fact that $Z = Y/X$, which would imply that $Z \sim 1/X$. $Z \sim 1/X$ is a very particular case of Menzerath-Altmann law that has been rejected by means of a correlation test between X and Y in genomes, being X the number of chromosomes of a species, Y its genome size in bases and Z the mean chromosome size. Here we provide rigorous statistical arguments to support the correctness of rejecting $Z \sim 1/X$ when X and Y are significantly correlated. Furthermore, we correct the recent claim that $Z \sim 1/X$ holds if and only if X and Y are independent (Baixeries et al. 2012, Biosystems 107 (3), 167-173). Indeed, $Z \sim 1/X$ if and only if Y is mean independent of X, a statistical property of intermediate strength between independence and uncorrelation. However, it is still true that the random breakage model proposed to explain $Z \sim 1/X$ with independence between X and Y does not fit real genome data. We reject $Z \sim 1/X$ in ten out of eleven taxonomic groups by means of a new correlation ratio test.

[*] Corresponding author.

Email addresses: `rferrericancho@lsi.upc.edu` (Ramon Ferrer-i-Cancho), `jbaixer@lsi.upc.edu` (Jaume Baixeries), `antonio.hernandez@upc.edu` (Antoni Hernández-Fernández)

Preprint submitted to Elsevier *October 25, 2012*

Keywords: Menzerath-Altmann law, scaling laws, genomes, language, independence, mean independence, uncorrelation

1. Introduction

Consider that X and Y are two discrete random variables and that $Z = Y/X$ with $X \neq 0$. For the particular case that Z is a mean, Menzerath's law is the tendency of Z to decrease as X increases: X stands for the number of parts of the construct (e.g., the number of clauses of a sentence), Y stands for the size of the whole in parts (e.g., the length in words of the sentence) and Z stands for the mean size of the construct (e.g., the mean length of the clauses in words). The dependency between Z and X that Menzerath's law describes qualitatively is typically modelled by means of Menzerath-Altmann law (Altmann, 1980), i.e.

$$Z = aX^b e^{cX}, \tag{1}$$

where a, b and c are constants. Menzerath's law has been found in language (Menzerath, 1954) and genomes (Ferrer-i-Cancho and Forns, 2009; Li, 2012) and also through a wide range of studies where Menzerath-Altmann law is fitted to human language (e.g., Altmann, 1980; Teupenhayn and Altmann, 1984), and music (Boroda and Altmann, 1991) and genomes (Wilde and Schwibbe, 1989; Li, 2012) which yields constants a, b and c that support a negative correlation between Z and X at least for sufficiently large X (see Cramer (2005) for a review of parameter values). Recently, it has been argued that $Z = Y/X$ leads inevitably to (Solé, 2010)

$$Z = aX^{-1}, \tag{2}$$

that is Menzerath-Altmann law with $b = -1$ and $c = 0$. If the argument was correct, Menzerath-Altmann law would be a trivial law at least from a mathematical perspective. Whether Eq. 2 holds in genomes, being X the number of chromosome of a species, Y the genome size in bases of that species and Z its mean chromosome size, has been debated (Solé, 2010; Ferrer-i-Cancho et al., 2012; Hernández-Fernández et al., 2011; Baixeries et al., 2012a,b). Here we aim to provide some theoretical foundations for testing if Eq. 2 holds and correct some incorrect mathematical arguments of Baixeries et al. (2012a). In particular, it will be shown that rejecting Eq. 2 if X and Y are correlated (Baixeries et al., 2012a; Hernández-Fernández et al., 2011) is correct but it will be shown that the claim that Eq. 2 is equivalent to independence between X and Y (Baixeries et al., 2012a) is incorrect because that equation holds if and only if Y is mean independent of X. Furthermore, a new test that rejects Eq. 2 in all taxonomic groups considered so far except one will be presented.

2. Mathematical preliminaries

2.1. The meaning of $Z = a/X$.

According to standard modelling, claiming that Eq. 2 holds can be recast as (Ritz and Streibig, 2008, pp. 1),

$$E[Z|X = x] = a/x, \tag{3}$$

for any x, being $E[Z|X = x]$ the conditional expectation of Z given x (a concrete value of X).

The next Lemma indicates that testing if a real sample follows $E[Z|X = x] = a/x$ is equivalent to testing if Y is mean independent of X.

Lemma 1. *Consider a constant a, two random natural variables, X and Y, and a third random number Z, such that $X > 0$ and $Z = Y/X$. Then, $E[Z|X = x] = a/x$ if and only if $E[Y|X = x] = a$ for any x.*

Proof. $E[Z|X = x] = E[(1/x)Y|X = x] = E[Y|X = x]/x$. □

So far, we have not cared about the value of a. The following theorem indicates that a is not arbitrary, showing the equivalence between the constancy of the conditional expectation and mean independence (Poirier, 1995, pp. 67), a condition that is well-known in econometrics (Cameron et al., 2009; Wooldridge, 2010).

Lemma 2. *Consider a constant a and two random natural variables, X and Y. $E[Y|X = x] = a$ for any x if and only if Y is mean independent of X, i.e. $E[Y|X = x] = E[Y]$ for any x.*

Proof. Showing that $E[Y|X = x] = E[Y]$ for any x implies $E[Y|X = x] = a$ for any x is trivial because $E[Y]$ does not depend on x. Now we aim to show that $E[Y|X = x] = a$ for any x implies $E[Y|X = x] = E[Y]$ for any x. By the law of total probability for expectations (DeGroot and Schervish, 2012, pp. 258), $E[Y] = E[E[Y|X = x]]$ and the substitution $E[Y|X = x] = a$, $E[Y] = E[a] = a$. □

The next theorem indicates that testing if Eq. 2 holds reduces to testing if $E[Z|X = x] = E[Y]$ for any x.

Theorem 3. *Consider a constant a and two random natural variables, X and Y, and a third random number Z, such that $X > 0$ and $Z = Y/X$. Then, $E[Z|X = x] = a/x$ if and only if Y is mean independent of X, i.e. $E[Y|X = x] = E[Y]$ for any x.*

Proof. Chaining Lemma 1 and Lemma 2, we have $E[Z|X = x] = a/x$ if and only if $E[Z|X = x] = E[Y]$ and by Lemma 2 we have $a = E[Y]$. □

We have examined one condition for a trivial Menzerath-Altmann law, namely that Y is mean independent of X. Another more obvious mathematically trivial version of the law occurs when Z is mean independent of X, i.e.

$$E[Z|X = x] = E[Z], \tag{4}$$

which is equivalent to constant $E[Z|X = x]$ by Lemma 2. The analysis of the correlation between Z and X in genomes discarded this constancy of Z for nine out of eleven taxonomic groups (Ferrer-i-Cancho and Forns, 2009, confirmed with an updated dataset by Baixeries et al. (2012a)). Therefore, Menzerath-Altmann law as a model of $E[Z|X = x]$ has two trivial versions:

- $b = c = 0$: Z is mean independent of X.

- $b = -1$ and $c = 0$: Y is mean independent of X.

Interestingly, b lies between 0 and -1 when $c = 0$ is assumed: e.g., $b = -0.27 \pm 0.11$ in language, being Z is the mean clause length in sentences and X is the number of sentences (Teupenhayn and Altmann, 1984), and $b = -0.44 \pm 0.09$ in music, being Z is the mean F-motif length in tones and X is the number of F-motifs (Boroda and Altmann, 1991). In both cases, we report $b = \mu \pm \sigma$, where b is the exponent of a sample while μ and σ are, the mean and the standard deviation of b in an ensemble of samples, respectively.

2.2. Three definitions of lack of association between X and Y.

For the remainder of sections, it is important to bear in mind the definition of three statistical properties between X and Y (Poirier, 1995, pp. 67-68):

- *X and Y are independent:* $p(Y = y|X = x) = p(Y = y)$ for any x and y.
- *Y is mean independent of X:* $E[Y|X = x] = E[Y]$ for any x.
- *X and Y are uncorrelated:* $COV(X, Y) = 0$ where $COV(X, Y) = E[XY] - E[X]E[Y]$ is the covariance between X and Y. Notice that uncorrelation, i.e. $\rho(X, Y) = 0$, being $\rho(X, Y)$ the Pearson correlation coefficient, is equivalent to zero covariance as (DeGroot and Schervish, 2012)
$$\rho(X, Y) = \frac{COV(X, Y)}{\sigma(X)\sigma(Y)}, \tag{5}$$
with $\sigma(X)$ and $\sigma(Y)$ as the standard deviation of X and Y, respectively.

As X and Y are uncorrelated if and only if $\rho(X, Y) = 0$ (or $COV(X, Y) = 0$), Y is mean independent of X if and only if $\eta(Y, X) = 0$, where $\eta(Y, X)$ is a less-known association metric: the correlation ratio Kruskal (1958). $\eta(Y, X)$ derives from the variance of $E[Y|X = X]$, which is by definition,

$$V[E[Y|X = x]] = E[(E[Y|X = x] - E[E[Y|X = x]])^2]. \tag{6}$$

By the law of total probability for expectations (DeGroot and Schervish, 2012, pp. 258), $E[E[Y|X = x]] = E[Y]$ and thus

$$V[E[Y|X = x]] = E[(E[Y|X = x] - E[Y])^2]. \tag{7}$$

From this variance, the correlation ratio of Y on X is defined as Kruskal (1958)

$$\eta(Y,X) = \left(\frac{V[E[Y|X=x]]}{V[Y]}\right)^{1/2}, \tag{8}$$

$$= \frac{\sigma[E[Y|X=x]]}{\sigma[Y]}, \tag{9}$$

where $\sigma(...)$ indicates the standard deviation. Notice that $0 \leq \eta(Y,X) \leq 1$ (whereas $-1 \leq \rho(X,Y) \leq 1$) (Kruskal, 1958, pp. 816-817). As $\rho(X,Y)$ is a normalized $COV(X,Y)$, $\eta(Y,X)$ is a normalized $V[E[Y|X=x]]$.

It is well-known that (Kolmogorov, 2008; Poirier, 1995):

$$\begin{gathered} X \text{ and } Y \text{ are independent} \\ \Downarrow \\ Y \text{ is mean independent of } X \quad (\eta(Y,X)=0) \\ \Downarrow \\ X \text{ and } Y \text{ are uncorrelated} \quad (\rho(X,Y)=0) \end{gathered}$$

Mean independence implies uncorrelation but the reverse (uncorrelation implies mean independence) is not true in general (see Appendix A for a counterexample). Similarly, independence implies mean independence but the reverse (mean independence implies independence) is not true in general (see Appendix A for a counterexample). Proofs of the top to bottom implications have been provided by, e.g., Kolmogorov (2008, pp. 60) or Poirier (1995, pp. 67).

In the next section it will be shown that the correlation ratio is indeed more powerful than a correlation coefficient for testing if Eq. 2 holds.

3. How to test that $Z = a/X$.

3.1. *A powerful test to reject $Z = a/X$.*

For the analyses of this section, we used the same dataset used by Hernández-Fernández et al. (2011); Baixeries et al. (2012a,b); Ferrer-i-Cancho et al. (2012). As the fact that Y is mean independent of X implies uncorrelation, i.e. $COV(X,Y) = 0$, Eq. 2 can be tested by means of the following procedure: if $COV(X,Y)$ is significantly different from 0 then reject Eq. 2, otherwise accept it. That procedure

Taxonomic group	N	$\rho(X,Y)$	p-value
Fungi	56	0.41	0.002
Angiosperms	4706	-0.0024	0.9
Gymnosperms	170	0.1	0.2
Insects	269	0.09	0.1
Reptiles	170	0.31	10^{-5}
Birds	99	-0.029	0.8
Mammals	371	0.3	$< 10^{-5}$
Cartilaginous fishes	52	0.014	0.9
Jawless fishes	13	-0.76	0.003
Ray-finned fishes	647	0.47	$< 10^{-5}$
Amphibians	315	0.13	0.02

Table 1: Analysis of the association between Y (genome size in bases) and X (chromosome number) by means of a Pearson correlation test. N is the number of species (the sample size), $\rho(X,Y)$ is the sample Pearson correlation coefficient and p-value is an estimation of the probability that a permutation of X yields an association at least as large as $|\rho(X,Y)|$ (the test is two-sided). Values of $\rho(X,Y)$ and p-value were rounded to leave only two or only one significant digit, respectively. $R = 10^5$ permutations were used.

can be used to reject Eq. 2 in genomes, being Z the mean chromosome length in bases of a species and X being the number of chromosomes of that species (Hernández-Fernández et al., 2011) Table 1 summarizes the results of the analysis of the Pearson correlation between X (chromosome number) and Y (genome size in bases) in genomes using a permutation test (a particular case of randomization test (Sokal and Rohlf, 1995, pp. 803-819)). A significant correlation is found in six out of eleven taxonomic groups. The test is conservative as it rejects Eq. 2 indirectly by means of a necessary condition for this equation to hold: $COV(X,Y) = 0$. Therefore, the five two groups where the Eq. 2 could not be rejected might be false negatives. Pearson correlation is a measure of linearity and has difficulties for capturing non-linear dependencies. A possible improvement is using a more powerful correlation metric such as $\rho_S(X,Y)$, the Spearman rank correlation coefficient, which is a measure of monotonic (linear or non-linear) dependency Zhou et al. (2003). The Spearman rank correlation test revealed that the majority of taxonomic groups (nine out of eleven) exhibit a significant correlation between X and Y that is incompatible with Eq. 2 (Hernández-Fernández et al., 2011). The exceptions are birds and cartilaginous

Taxonomic group	N	$\eta(Y, X)$	p-value
Fungi	56	0.74	0.03
Angiosperms	4706	0.27	0.001
Gymnosperms	170	0.74	$< 10^{-5}$
Insects	269	0.46	0.02
Reptiles	170	0.48	0.003
Birds	99	0.78	0.001
Mammals	371	0.68	$< 10^{-5}$
Cartilaginous fishes	52	0.76	0.1
Jawless fishes	13	0.98	0.05
Ray-finned fishes	647	0.69	$< 10^{-5}$
Amphibians	315	0.58	$< 10^{-5}$

Table 2: Analysis of the association between Y (genome size in bases) and X (chromosome number) by means of a correlation ratio test. The format is similar to that of Table 1. $\eta(Y, X)$ is the sample correlation ratio of Y on X, p-value is an estimation of the probability that a permutation of X yields an association as large as $\eta(Y, X)$ (the test is one-sided). $R = 10^5$ permutations were used.

fishes. However, there is a more powerful way of testing Eq. 2: testing directly if Y is mean independent of X from its definition, i.e. $E[Y|X = x] = E[Y]$ for any x, which is equivalent to $\eta(X, Y = 0$ (Kruskal, 1958).

Table 2 summarizes the results of the analysis of the correlation ratio of Y (genome size in bases) on X (chromosome number) in genomes using a permutation test. $\eta(Y, X)$ was not significantly large in one taxonomic group: cartilaginous fishes. Mean independence and, equivalently, Eq. 2, cannot be rejected for only one out of eleven groups.

Table 3 summarizes the results of the Pearson, Spearman and correlation ration test. The number of taxonomic groups for which a test rejects Eq. 2 is 6, 9, and 10, respectively. The qualitative summary for $\rho(X, Y)$ and $\eta(Y, X)$ comes, respectively, from Table 1 and 2. The qualitative summary for $\rho_S(X, Y)$ is due to Hernández-Fernández et al. (2011). The power of the correlation ratio test can be easily explained by the fact that it can only be zero when mean independence fails. Table 3 suggests that Spearman rank correlation has an intermediate power between Pearson correlation and correlation ratio.

Taxonomic group	$\rho(X,Y)$	$\rho_S(X,Y)$	$\eta(Y,X)$
Fungi	Yes	Yes	Yes
Angiosperms		Yes	Yes
Gymnosperms		Yes	Yes
Insects		Yes	Yes
Reptiles	Yes	Yes	Yes
Birds			Yes
Mammals	Yes	Yes	Yes
Cartilaginous fishes			
Jawless fishes	Yes	Yes	Yes
Ray-finned fishes	Yes	Yes	Yes
Amphibians	Yes	Yes	Yes

Table 3: Summary of the analysis of the association between X (genome size in bases) and Y (chromosome number) in genomes. Three statistics are considered: the sample Pearson correlation coefficient $\rho(X,Y)$, the sample Spearman correlation coefficient $\rho_S(X,Y)$ and the sample correlation ratio ($\eta(Y,X)$). 'Yes' indicates that the corresponding correlation test indicates a significant correlation at a significance level of 0.05.

3.2. A weak test to accept $Z = a/X$.

The hypothesis of $Z = a/X$ has been accepted with the only support that the fit of $Z = aX^b$ yields $b \approx -1$ (Solé, 2010). This procedure is very prone to type II error (accepting a false null hypothesis) as it needs that $Z = aX^b$ holds first (Baixeries et al., 2012b; Ferrer-i-Cancho et al., 2012). Our analysis shows that for $Z = a/X$ to hold, it is not only necessary that $b \approx -1$ is retrieved but also $a \approx \mu[Y]$, where $\mu[Y]$ is the mean of Y, an estimate of $E[Y]$ (recall Theorem 3). Even if $a \approx \mu[Y]$ and $b \approx -1$, type II errors are not excluded and minimizing them needs evidence that $Z = aX^b$ is well-supported by data.

4. Theorem 1 and Corollary 1 by Baixeries et al. are incorrect

In a recent paper, if has been claimed that (Theorem 1 by Baixeries et al. (2012a)) that

Two random natural numbers X and Y, such that $X > 0$, are independent if and only if $Z = Y/X$ satisfies $E[Z|X] = E[Y]/X$.

In order to be correct, that theorem should state

Given *two random natural* **variables** X *and* Y*, such that* $X > 0$*,* Y **is mean independent of** X *if and only if* $Z = Y/X$ *satisfies* $E[Z|X] = E[Y]/X$*.*

or

If *two random natural* **variables** X *and* Y*, such that* $X > 0$*, are independent* **then** $Z = Y/X$ *satisfies* $E[Z|X] = E[Y]/X$*.*

The point is Theorem 3 and the fact that independence implies mean independence (Kolmogorov, 2008; Poirier, 1995) but the reverse does not hold (Appendix A). The proof provided by Baixeries et al. (2012a) for their incorrect theorem indeed only demonstrates that independence implies $E[Z|X] = E[Y]/X$. Theorem 3 and the fact that independence is more restrictive than mean independence (Appendix A) indicate that the converse is impossible to prove.

Baixeries et al. (2012a) studied a general class of random models where X and Y are independent, with $Z = Y/X$, which is generalization of a random model where X and Y are independent but distributed uniformly (Solé, 2010). For the generalized class where X and Y remain independent but not necessarily distributed uniformly, the following corollary was presented (Corollary 1 by Baixeries et al. (2012a)):

The general class of random models above is equivalent to the class of models yielding $E[L_c|L_g] = aL_g^{-\beta}$*, with* $a = E[G]$ *and* $\beta = 1$*.*

(the meaning of the notation is the following: $X = L_g$, $Y = G$, $Z = L_c = G/L_g$ and $b = \beta$).

That corollary is incorrect as these random models are only a subclass of all the models yielding Eq. 2. In order to be correct, it should state

The general class of random models above **is a strict subset of** *the class of models yielding* $E[L_c|L_g] = aL_g^{-\beta}$*, with* $a = E[G]$ *and* $\beta = 1$*.*

Again, the point is that independence implies mean independence but not the reverse. Therefore, the class of generalized class of random models where X and Y are independent covers only a fraction of the models that according to Theorem 3 lead to Eq. 2. Further corrections needed by Baixeries et al. (2012a) are provided in Appendix B.

5. Discussion

We have shown that

- Claiming that Z and X follow a very particular form of Menzerath-Altmann law, i.e. Eq. 2, is indeed equivalent to claiming that
$$E[Z|X] = E[Y]/X. \tag{10}$$
holds.
- Claiming that Eq. 10 holds is equivalent to claiming that Y is mean independent of X.
- The previous claim that independence between X and Y is equivalent to mean independence (Baixeries et al., 2012a) is wrong.
- A new correlation ratio test shows that Eq. 10 could only hold in cartilaginous fishes.

It is still true that the random breakage model where X and Y are independent and uniformly distributed (Solé, 2010) fails to fit the majority of taxonomic groups because independence is a particular case of mean independence and mean independence needs fails in at least ten out of eleven taxonomic groups (Table 3). Furthermore, Baixeries et al. (2012a) have argued that independence between X and Y is problematic as it can lead to organisms with empty chromosomes or empty chromosome parts. Interestingly, Y does not need to be the size of genome in bases. It could be the size in units between the base and the chromosome.

The problem of empty components also concerns mean independence. The condition for not expecting empty chromosomes for a given x (a concrete value of X) is

$$\begin{aligned} E[Z|X=x] &\geq 1 \\ \frac{1}{x}E[Y|X=x] &\geq 1. \\ E[Y|X=x] &\geq x. \end{aligned} \tag{11}$$

For that x, the condition in Eq. 11 becomes $E[Y] \geq x$ when Y is mean independent of X as $E[Y|X=x] = E[Y]$ in that case. Thus, under mean independence, empty chromosomes are expected in an organisms of x chromosomes if $E[Y] < x$. Notice that expecting that Eq. 11 holds on average for any x, leads to

$$\begin{aligned} E[E[Y|X=x]] &\geq E[X] \\ E[Y] &\geq E[X] \end{aligned} \tag{12}$$

thanks to the law of total probability for expectations (DeGroot and Schervish, 2012, pp. 258). The restrictions defined by Eqs. 11 and 12 are perhaps very simple but Baixeries et al. (2012a) considered more elaborated constraints for the viability of an organism based upon the parts making an ideal chromosome: a centromere, two telomeres and a couple of intermediate regions. Those viability constraints lead to a deviation from Menzerath-Altman law with $b = -1$ and $c = 0$ (see Fig. 3 of Baixeries et al. (2012a)), which Theorem 3 allows one to interpret unequivocally as a departure from mean independence. Therefore, the viability and well-formedness of chromosomes is not compatible with mean independence either. It is still true that the negative correlation between Z and X, known as Menzerath's law, defies a trivial explanation in genomes (Ferrer-i-Cancho and Forns, 2009; Wilde and Schwibbe, 1989). Claiming that Eq. 2 is inevitable (Solé, 2010) is equivalent to claiming that Y must be mean independent of X in any circumstance, a very strong requirement for real language, music and genomes.

Acknowledgements

We are grateful to Ł. Debowski, P. Delicado, R. Gavaldà and J. Mačutek for their valuable mathematical insights. We owe the counterexample of Appendix A.2 to P. Delicado. We are also grateful to G. Bel-Enguix and N. Forns for helful discussions and to E. Pons for his advice. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

References

Altmann, G., 1980. Prolegomena to Menzerath's law. Glottometrika 2 2, 1–10.

Baixeries, J., Hernández-Fernández, A., Ferrer-i-Cancho, R., 2012a. Random models of Menzerath-Altmann law in genomes. Biosystems 107, 167173.

Baixeries, J., Hernández-Fernández, A., Forns, N., Ferrer-i-Cancho, R., 2012b. The parameters of Menzerath-Altmann law in genomes. Journal of Quantitative Linguistics.
URL http://arxiv.org/abs/1201.1746

Boroda, M. G., Altmann, G., 1991. Menzerath's law in musical texts. Musikometrica 3, 1–13.

Cameron, A., , Trivedi, P. K., 2009. Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge.

Cramer, I. M., 2005. The parameters of the Altmann-Menzerath law. Journal of Quantitative Linguistics 12 (1), 41–52.

DeGroot, M. H., Schervish, M. J., 2012. Probability and statistics, 4th Edition. Wiley, Boston.

Ferrer-i-Cancho, R., Forns, N., 2009. The self-organization of genomes. Complexity 15 (5), 34–36.

Ferrer-i-Cancho, R., Forns, N., Hernández-Fernández, A., Bel-Enguix, G., Baixeries, J., 2012. The challenges of statistical patterns of language: the case of Menzerath's law in genomes. Complexity, in press.

Hernández-Fernández, A., Baixeries, J., Forns, N., Ferrer-i-Cancho, R., 2011. Size of the whole versus number of parts in genomes. Entropy 13, 1465–1480.

Kolmogorov, A. N., 2008. Theory of probability, 2nd Edition. Chelsea Publising Company, New York.

Kruskal, W. H., 1958. Ordinal measures of association. Journal of the American Statistical Association 53, 814–861.

Li, W., 2012. Menzerath's law at the gene-exon level in the human genome. Complexity 17 (4), 49–53.

Menzerath, P., 1954. Die Architektonik des deutschen Wortschatzes. Dümmler, Bonn.

Poirier, D. J., 1995. Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press, Cambridge.

Ritz, C., Streibig, J. C., 2008. Nonlinear regression with R. Springer, New York.

Sokal, R. R., Rohlf, F. J., 1995. Biometry. The principles and practice of statistics in biological research, 3rd Edition. W. H. Freeman and Co., New York.

Solé, R. V., 2010. Genome size, self-organization and DNA's dark matter. Complexity 16 (1), 20–23.

Teupenhayn, R., Altmann, G., 1984. Clause length and menzerath's law. Glottometrika 6, 127–138.

Wilde, J., Schwibbe, H., 1989. Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. In: Altmann, G., Schwibbe, M. H. (Eds.), Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Olms, Hildesheim, pp. 92–107.

Wooldridge, J. M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge.

Zhou, K., Tuncali, K., Silverman, S. G., 2003. Correlation and simpler linear regression. Radiology 227, 617–628.

Appendix A. Counterexamples

Appendix A.1. If Y is mean independent of X then X and Y are not necessarily independent

Consider that

$$p(X = x, Y = y) = \begin{cases} 1/2 & \text{if } x = 0 \text{ and } y = 0 \\ 1/4 & \text{if } x = 1 \text{ and } y = 1 \\ 1/4 & \text{if } x = 1 \text{ and } y = -1 \end{cases} \tag{A.1}$$

Thus $E(Y|X = 1) = (1/2)(-1) + (1/2)1 = 0$ and $E(Y|X = 0) = 0$. Therefore Y is mean independent of X but X and Y are not independent because $p(X = 1, Y = 0) = 0 \neq p(X = 1)p(Y = 0)$ as $p(X = 1) = 1/2$ and $p(Y = 0) = 1/2$.

Appendix A.2. Uncorrelation between X and Y does not imply that Y is mean independent of X

Consider that

$$p(X = x, Y = y) = \begin{cases} 1/2 & \text{if } x = 0 \text{ and } y = -1 \\ 1/4 & \text{if } x = -1 \text{ and } y = 1 \\ 1/4 & \text{if } x = 1 \text{ and } y = 1 \end{cases} \tag{A.2}$$

Thus $E(X) = E(Y) = 0$ and $E(XY) = (1/4)(-1) + (1/2)0 + (1/4)1 = 0$. Therefore $COV(X, Y) = 0$ but Y is not mean independent of X because $E(Y|X = -1) = 1 \neq E(Y|X = 0) = -1$.

Appendix B. Corrections on Baixeries et al. (2012a)

Besides the theorem and the corollary revised in Section 4, Baixeries et al. (2012a) needs corrections in other places:

- p.168 *Menzerath-Altmann law with* $b = 1$ *and* $c = 0$ *is equivalent to independence between* G *and* L_g.

  should read

  Menzerath-Altmann law with $b = 1$ *and* $c = 0$ *is equivalent to* G **being mean independent of** L_g.

- p. 168 *the need of independence between* G *and* L_g *to obtain Menzerath-Altmann law with* $b = 1$ *and* $c = 0$.

  should read

  the need **that** G **is mean independent of** L_g *to obtain Menzerath-Altmann law with* $b = 1$ *and* $c = 0$.

- p.170 *In all the calculations that follow, true independence between* G *and* L_g *is assumed.*

  should read

  In all the calculations that follow, **mean independence between** G **and** L_g **(which includes independence between** G **and** L_g **as a particular case)** *is assumed.*

- p.171. *Corollary 1 states that* $E[L_c|L_g] \sim 1/L_g$ *can only be achieved by independence between* G *and* L_g

  should read

  Corollary 1 states that $E[L_c|L_g] \sim 1/L_g$ *can only be achieved* **if** G **is mean independent of** L_g.